The *BioLexicon*: a Large-Scale Domain-Specific Lexical Resource for Biomedical Text Mining

Simonetta Montemagni

Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR

simonetta.montemagni@ilc.cnr.it

LREC 2010

2nd Workshop on **Building and evaluating resources for biomedical text mining**

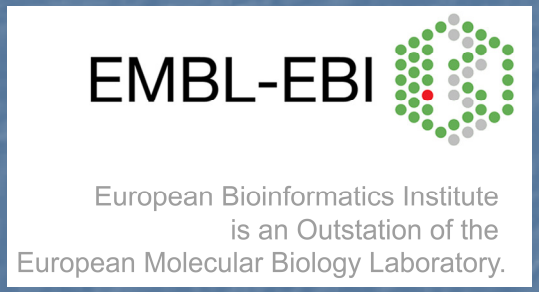
Malta – May 18th

Outline

- The BioLexicon
  - Who
  - Why
  - How
  - What
  - Where from
    - Focus on Verbs
  - Representation
  - How many entries
  - Evaluation
  - Distribution
  - Conclusions

The BioLexicon: **who**

Joint and collaborative work of the following teams in the framework of the European **BOOTStrep** project (FP6 - 028099)



- **D. Rebholz-Schuhmann, P. Pezik, V. Lee, J.J. Kim**
  European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK

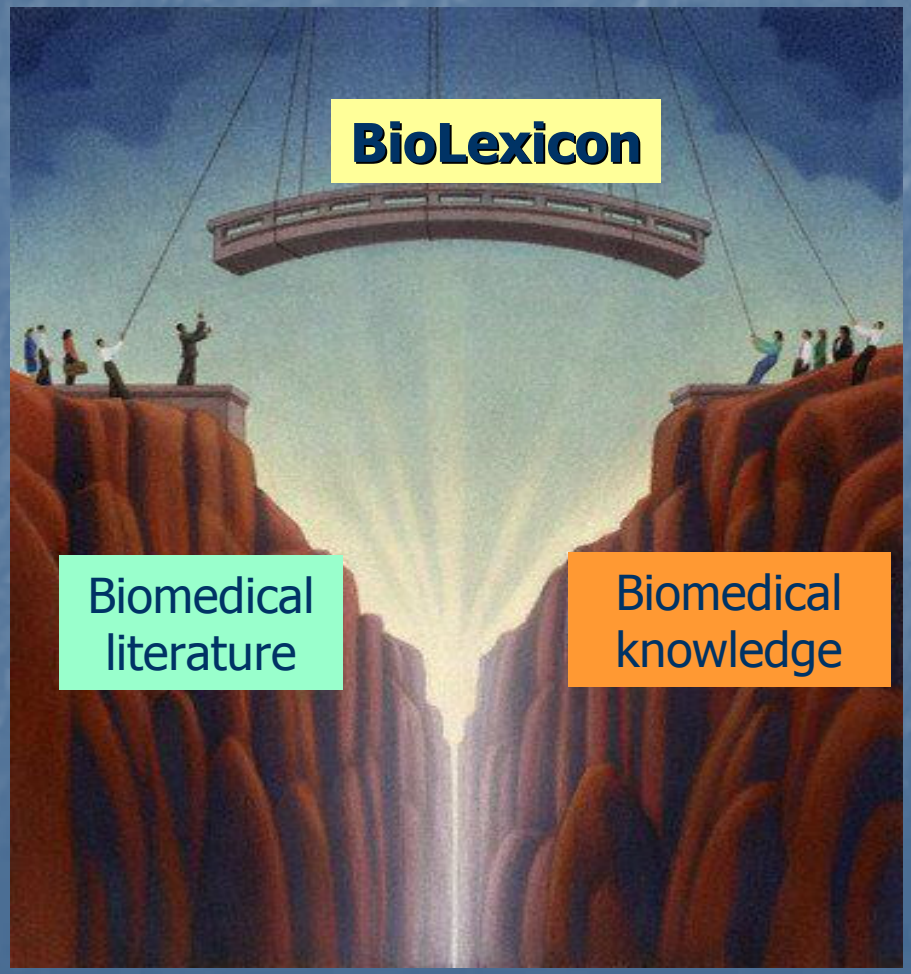
- **N. Calzolari, M. Monachini, S. Montemagni, R. del Gratta, S. Marchi, V. Quochi, G. Venturi**
  Istituo di LInguistica Computazionale "Antonio Zampolli" - CNR, Via Giuseppe Moruzzi N° 1, 56124 Pisa, Italy

- **S. Ananiadou, J. McNaught, Y. Sasaki, P. Thompson**
  School of Computer Science, The University of Manchester, 131 Princess Street, M1 7DN, UK

The BioLexicon: **why**

- Text Mining needs information about words
  - the lexical component still remains a major bottleneck
- TM systems in the biomedical domain must be provided with a substantial lexicon covering a realistic vocabulary and providing the kinds of linguistic information appropriate to grasp the knowledge embedded in texts
  - Biomedical term variants (orthographic, semantic, geographical, …)
    - better information retrieval
  - Terminological verbs and their combinatorial properties (subcategorization frames and predicate-argument structure)
    - better information extraction and question answering
  - Word derivations
    - to reach similar meaning expressed in different ways (e.g. *activation* vs *activate*)

The BioLexicon as a key ingredient to **bridge the gap** between text and knowledge



To perform such a key role the BioLexicon MUST

- reflect the actual usage of words in biomedical texts
- be continuously updated with new word synonyms emerging from texts
- include rich linguistic information on the behavioural properties of nouns and verbs

The BioLexicon: **how**

- **General Requirements**

  Modularity, extensibility, conformity to standards, reusability

- **Biomedical Domain Specific Requirements**

  Gene names, protein names, bio-events and participants, …

- **Linguistic/Terminological Requirements**

  term variants, source identifiers, acronyms, syntactic and semantic properties of terms, …

- **Text Mining / Machine Learning Requirements**

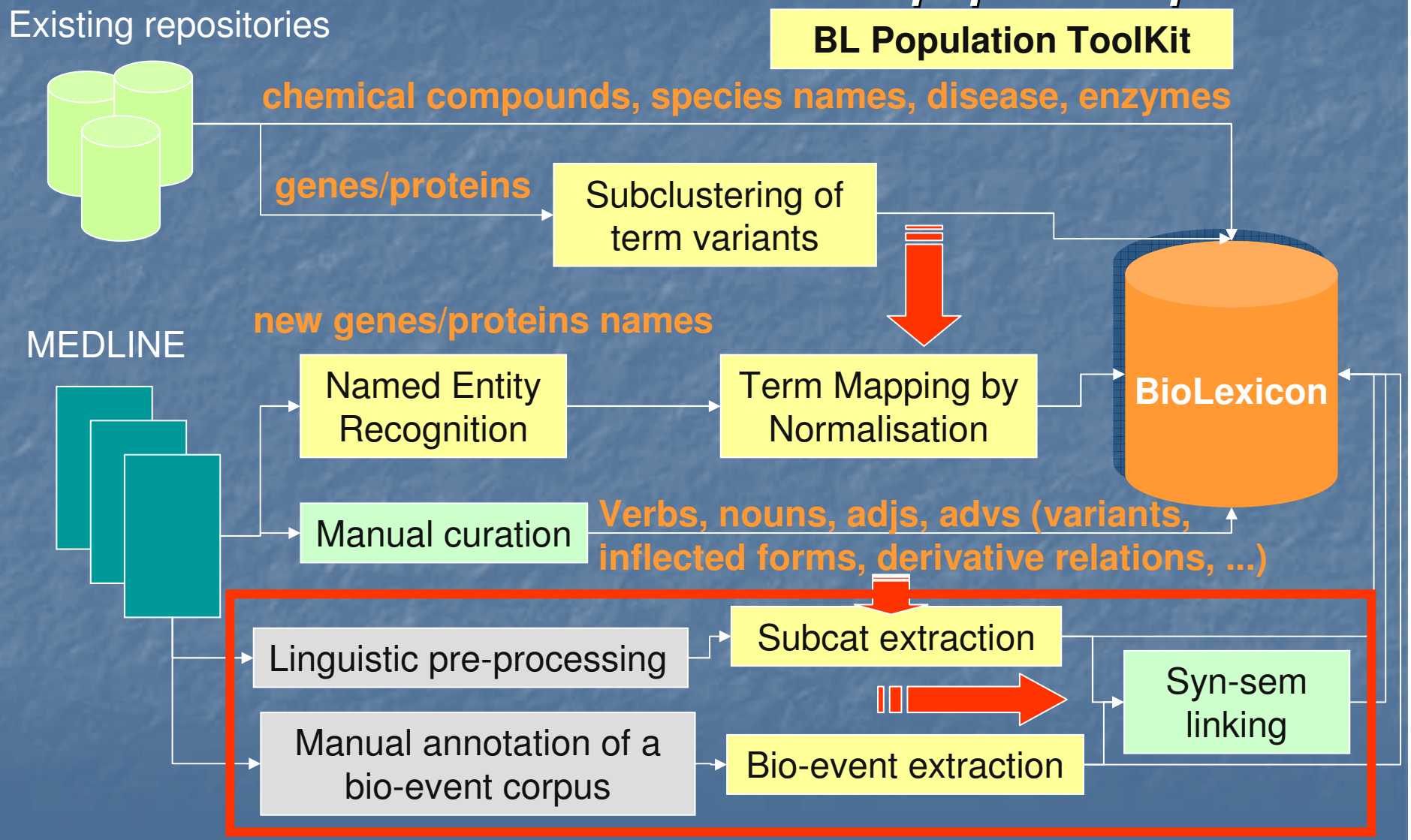
  Confidence scores for automatically extracted info (e.g. variants, subclusterizations, subcat frames, …)

The BioLexicon: **what**

- integrated lexical-terminological resource of ~2.2M lexical entries for bio-text mining with information about
  - nouns, verbs, adjectives, adverbs
  - both domain-specific and general language words
- populated with terms gathered from
  - available biomedical sources
  - texts (biomedical literature)
- including rich linguistic information ranging over different linguistic descriptions levels
  - e.g. derivational morphology, subcategorization patterns, predicate argument structure, syntax-semantics linking
- combining features of both terminologies and open-domain computational lexicons
- conforming to international lexical representation standards (the ISO/DIS 24613 "Lexical Mark-up Framework")
- providing links to the Gene Regulation Ontology

The BioLexicon: **where from**

Incremental population process



The BioLexicon: **focus on verbs**

- Accurate TM applications focused on event extraction require lexical resources providing an exhaustive account of the **semantic and syntactic combinatorial properties of lexical units** conveying **event information**
  - Several exist for the general language domain, e.g. FrameNet, VerbNet, PropBank
- Specialist domains such as *biology* require domain-specific resources
  - use of different predicates to describe events
    - E.g. *methylate, phosphorylate*
  - general language predicates may have different properties
    - E.g. *the patient **presented** with influenza to the doctor* vs *the patient **presented** the doctor with influenza*

Current biomedical lexical resources

- A number of attempts have been made to produce domain-specific extensions of general-purpose lexical semantic resources providing information on predicate-argument structure
  - BioFrameNet and PASBio
    - corpus-based
    - small-scale
  - SPECIALIST lexicon
    - extension of a large lexicon of general English
    - not corpus-driven
    - syntactic complementation patterns only
- Creation of resources focussed on predicate-argument structure can be a major bottleneck
  - Mostly manually created by lexicographers
  - Limited coverage
  - Time-consuming to port to new domains
- Automatic or semi-automatic acquisition methods more promising and increasingly viable
  - Advances in NLP and machine learning technology
  - Availability of corpora

Current biomedical lexical resources: the need

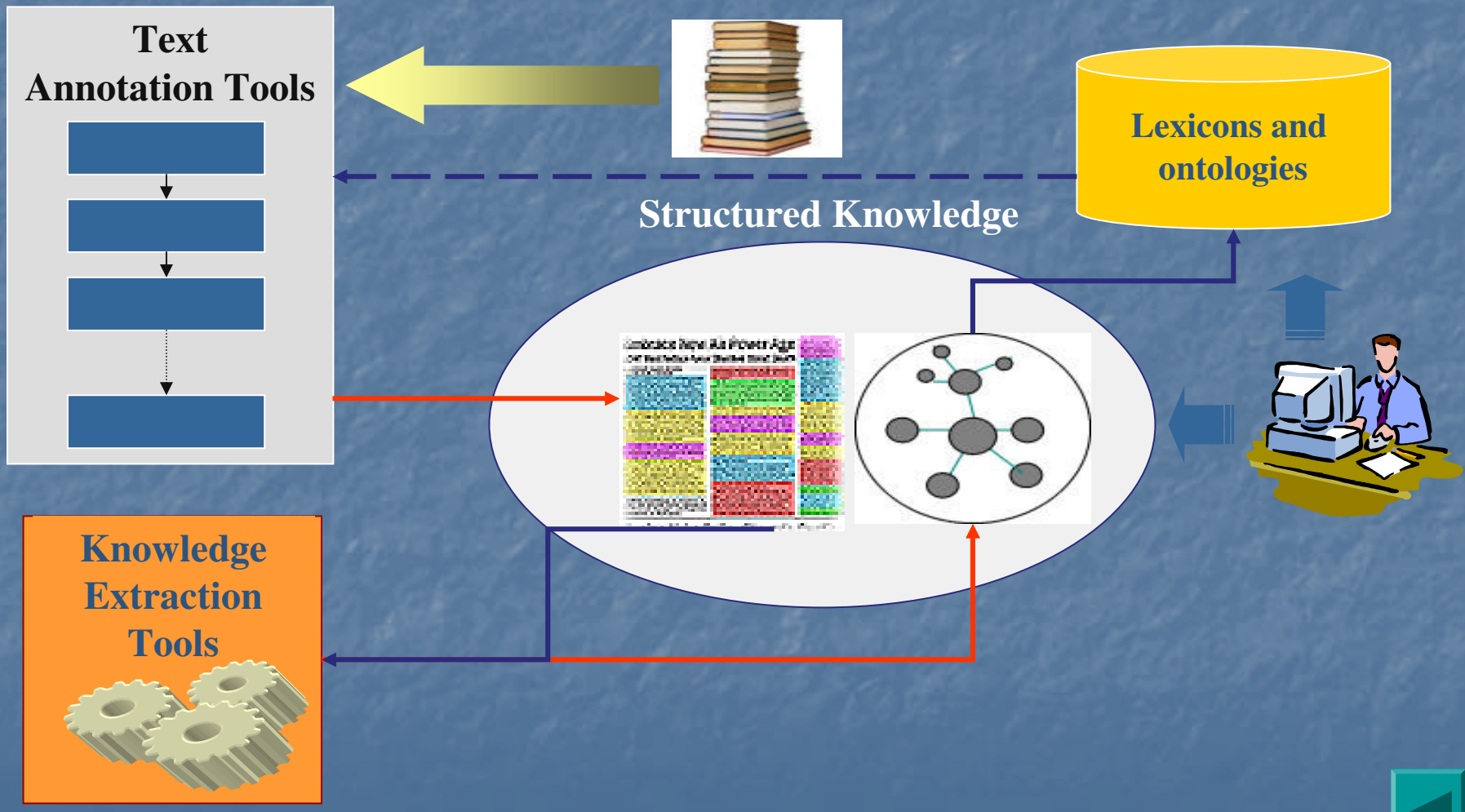
- To our knowlege, there is currently no large-scale domain-specific lexical resource providing predicate-argument information
  - based on domain-specific corpora
  - containing both syntactic and semantic information

The BioLexicon as a possible answer

The BioLexicon: our approach towards acquisition of **verb information**

- bootstrapped from biomedical corpora
  - the most relevant verbs are included in the lexicon
  - their encoded behaviour is domain-specific
- contains *both* syntactic and semantic information
  - syntactic subcategorization (e.g. *act* ARG1#PP-*as#*)
  - semantic event frame information (e.g. *bind* AGENT#THEME#LOC#)
  - explicit link between the two (e.g. *express* AGENT>ARG1#THEME>ARG2#LOC>PP-*in#*)
- built semi-automatically by combining NLP and Machine Learning techniques
  - syntactic frames are extracted through unsupervised learning on dependency-annotated text
  - semantic frames are based on manual annotation of gene regulation bio-events by domain experts
  - link between syntactic and semantic information is manually added

From Text to Knowledge: NLP and Knowledge Extraction



The BioLexicon: our approach to **subcat** extraction

- acquired through unsupervised automatic acquisition techniques from linguistically pre-processed domain corpora
  - the starting point: shallow or deep syntactic annotation?
- particular requirements for Subcategorization Frames (SCFs) in biomedical language
  - SCFs should also include strongly selected modifiers (such as location, manner and timing), as these are deemed to be essential for the correct interpretation of texts
  - average number of arguments in SCFs higher than general language
- "discovery" approach to SCF acquisition based on a looser notion of SCFs, which includes typical verb modifiers in addition to strongly selected arguments
  - no a priori knowledge about the set of possible SCFs
  - no distinction between argument/modifier
- **to meet this basic requirement, SCF induction operates on a deep level of syntactic annotation**

The BioLexicon: the **subcat** extraction process

- SCFs extracted from a corpus of MEDLINE abstracts and full papers made up of 6 million word tokens
- The induction process was performed through:
  - syntactic annotation of the acquisition corpus with Enju syntactic parser (v2.2, http://www-tsujii.is.s.u-tokyo.ac.jp/enju/) (adapted to biomedical texts)
  - extraction of the observed dependency sets (ODSs) for each verbal occurrence
    - each ODS represented as a set of dependencies described in terms of relation type (e.g. ARG1, ARG2, PP-in, PP-across, that-CL, etc.)
    - order of dependencies in each ODS is normalised
  - induction of relevant SCF associated with a given verb
    - for each ODS type, the conditional probability given the verb type *v* was computed
    - weighted thresholds used to filter out noisy frames
      - an ODS type with an associated probability score beyond a certain threshold is selected as eligible SCF for that verb type
    - each SCF has been extracted for one *normalised verb token*, i.e. the extraction process makes abstraction from the passive usages

*Overproduction of TorR in a torT strain **resulted** in partial constitutive expression of the torA'-'lacZ fusion, **suggesting** that TorR **acts** downstream from TorT.*

ROOT	ROOT	ROOT	-1	ROOT	resulted	result	VBD	7
resulted	result	VBD	7	verb_arg1	Overproduction	overproduction	NN	0
a	a	DT	4	det_arg1	strain	strain	NN	6
the	the	DT	13	det_arg1	fusion	fusion	NN	15
torT	tort	NN	5	noun_arg1	strain	strain	NN	6
partial	partial	JJ	9	adj_arg1	expression	expression	NN	11
torA'-'lacZ	tora'-'lacz	NN	14	noun_arg1	fusion	fusion	NN	15
constitutive	constitutive	JJ	10	adj_arg1	expression	expression	NN	11
acts	act	VBZ	20	verb_arg1	TorR	torr	NN	19
of	of	IN	1	prep_arg12	Overproduction	overproduction	NN	0

act ARG1=torr@NN 0 0 0 0 PP-from=tort@NNP 0 0 0 0 0 0 0 0 0 0 VBZ 0
result ARG1=overproduction@NN 0 0 0 0 PP-in=expression@NN 0 0 0 0 MOD=suggest@VBG 0 0 0 0
suggest ARG1=UNKNOWN@UNKNOWN 0 0 0 0 0 0 0 0 0 0 0 that-CL=act@VBZ 0 0 VBG 0

of	of	IN	12	prep_arg12	expression	expression	NN	11
of	of	IN	12	prep_arg12	fusion	fusion	NN	15
that	that	IN	18	comp_arg1	acts	act	VBZ	20
downstream	downstream	RB	21	adj_arg1	acts	act	VBZ	20
in	in	IN	8	prep_arg12	resulted	result	VBD	7
in	in	IN	8	prep_arg12	expression	expression	NN	11
from	from	IN	22	prep_arg12	acts	act	VBZ	20
from	from	IN	22	prep_arg12	TorT	tort	NNP	23
suggesting	suggest	VBG	17	verb_mod_arg12	resulted	result	VBD	7
suggesting	suggest	VBG	17	verb_mod_arg12	UNKNOWN	UNKNOWN	UNKNOWN	-1
suggesting	suggest	VBG	17	verb_mod_arg12	acts	act	VBZ	20

The BioLexicon: **subcat** extraction results

- 136 different induced SCFs
  - vs SPECIALIST Lexicon: very limited number of complementation patterns

Verb	SCF	P(subcat|v)	Pass
abolish	ARG1#ARG2#	0.8669767	0.1437768
abolish	ARG1#ARG2#MOD@VBG#	0.0390697	0.1904761
abolish	ARG1#ARG2#PP-in#	0.0939534	0.7029702
accumulate	ARG1#ARG2#	0.2940677	0.0403458
accumulate	ARG1#	0.4627118	0
accumulate	ARG1#ARG2#PP-in#	0.1084745	0.140625
accumulate	ARG1#PP-in#	0.1347457	0

The BioLexicon: acquired SCFs and strongly selected modifiers

- many of the strongly selected modifiers spread over different SCFs
  - radically underestimated role
- SCFs complemented with information about individual dependencies of verbs
  - typical verbal dependencies, corresponding to either arguments or strongly selected modifiers, detected through the ll association score
  - 44 induced dependency types

v	SCF	v_freq	SCF_freq	p(SCF|v)	% passive usages
methylate	ARG1#ARG2#	422	294	0.6967	0.1258
methylate	ARG1#ARG2#PP-in#				
methylate	ARG1#ARG2#PP-at#				

v	DEP	all_ dep	dep_ freq	p(dep|v)	ll	% passive usages
methylate	ARG2#	1406	410	0.2916	778.5146	0.25
methylate	PP-at#	1406	29	0.0206	57.9113	0.31
methylate	PP-in#	1406	45	0.0320	18.2749	0.60

The BioLexicon: an example of stored subcat information for the verb *acquire*

v	*SCF*	*p(SCF|v)*	*% pass*
acquire	ARG1#ARG2#	0.5461	0.1284
acquire	ARG1#ARG2#PP-in#	0.0886	0.0833
acquire	ARG1#ARG2#PP-from#	0.0406	0.1818
acquire	ARG1#ARG2#PP-by#	0.0406	0.0000
acquire	ARG1#ARG2#PP-during#	0.0295	0.3750

Full parsing

v	*DEP*	*ll*	*% pass*
acquire	ARG2#	579.96392	0.1512915
acquire	WH-when#	25.703417	0.1
acquire	PP-from#	22.716082	0.3333333
acquire	PP-by#	13.626654	0
acquire	PP-in#	13.416025	0.1666667

Preposition-based parsing

The BioLexicon: contained verb subcategorization information

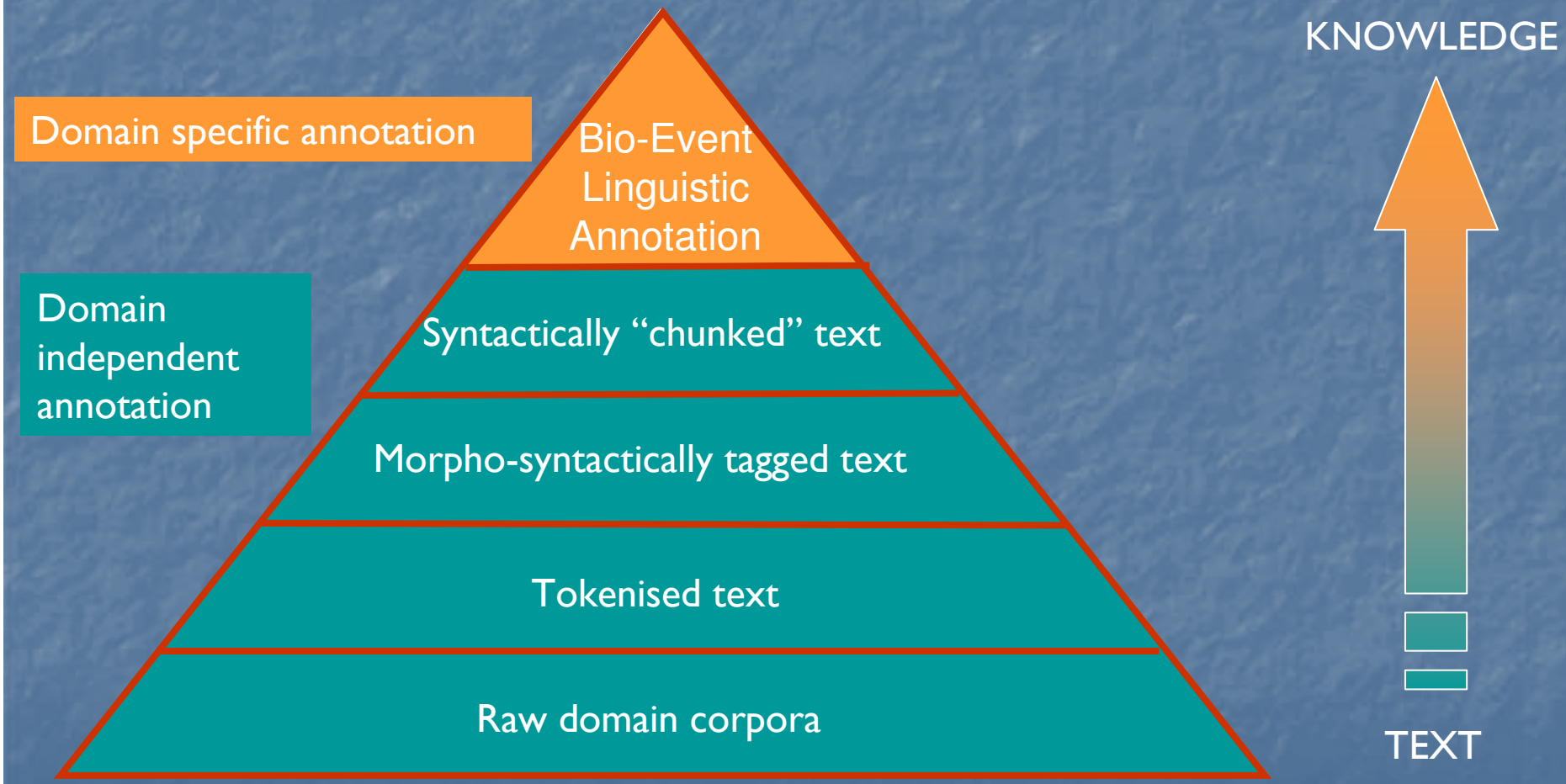
- complementarity between Verb-Dep and Verb-SCF associations
  - both information types included into the BioLexicon
- subcat frame information acquired for 759 different verbs, corresponding to 658 different base forms
  - e.g. the occurrences of *colocalize, colocalise*, *co-localize* and *co-localise* recorded under *colocalize*
- the BioLexicon was augmented with:
  - 1410 Verb-SCF associations, involving 136 different subcat frame types
  - 3040 Verb-Dep associations, involving 44 slot types

The BioLexicon: **bio-event frames**

- Extracted from a corpus of 677 MEDLINE abstracts manually annotated by biologists
  - semantic parsers not mature enough to provide the starting point (as opposed to dependency parsers)
  - manual semantic annotation carried out on top of shallow syntactic annotation ("chunking")
    - consistency of marked text spans helped by annotating syntactic chunks

[NP The narL gene product] [VP activates]
[NP the nitrate reductase operon] [PP in]
[NP Escherichia coli]

Bio-Event annotated corpus:
incremental annotation approach



The BioLexicon: **bio-event frames** annotation

- Annotation consisted of:
  - Identifying relevant *gene regulation* events centred on verbs and nominalised verbs (e.g. *expression*)
  - Finding all semantic arguments of an identified event
    - Within-sentence annotation
  - Assigning a semantic role to each argument
  - Assigning named entity types to semantic arguments (where appropriate)
    - A hierarchy of NEs, specially tuned to *gene regulation,* was created
    - Organised into five entity-specific super-classes

Thompson P., Cotter P., McNaught J., Ananiadou S., Montemagni S., Trabucco A., Venturi G. 2008. Building a Bio-Event Annotated Corpus for the Acquisition of Semantic Frames from Biomedical Corpora. In Proceedings of LREC 2008

Bio-Event annotated corpus:
semantic roles

- Aim for a set of **verb-specific event frames**
- Use of frame-independent semantic roles
  - annotation of **all sublanguage semantic arguments**, using a set of **domain-specific** and **domain-independent roles**
- The proposed set of 12 event-independent semantic roles includes:
  - two domain-specific semantic roles, i.e. CONDITION and MANNER;
  - semantic roles particularly important for the precise definition of complex biological relations, even though not necessarily specific to the field, i.e. LOCATION and TEMPORAL;
  - semantic roles widely traceable across all domains

Bio-Event annotated corpus:
list of semantic roles (1)

AGENT	Drives/instigates event	**The narL gene product** ***activates*** the nitrate reductase operon
THEME	a) Affected by/results from event b) Subject of events describing states	**recA protein** was ***induced*** by UV radiation **The FNR protein** *resembles* CRP
MANNER	Method/way in which event is carried out	cpxA gene ***increases*** the levels of csgA transcription by **dephosphorylation** of CpxR
INSTRUMENT	Used to carry out event	We have ***isolated*** a strain with the aid of **the Casadaban Mud phage**
LOCATION	Where *complete* event takes place	Phosphorylation of OmpR ***modulates*** expression of the ompF and ompC genes **in Escherichia coli**
SOURCE	Start point of event	A transducing lambda phage was ***isolated*** from **a strain** harboring a glpD''lacZ fusion
DESTINATION	End point of event	Transcription of gntT is activated by ***binding*** of the cyclic AMP (cAMP)-cAMP receptor protein (CRP) complex to **a CRP binding site**

Bio-Event annotated corpus:
list of semantic roles (2)

TEMPORAL	Situates event w.r.t another event	The Alp protease activity is ***detected*** in cells **after introduction** of plasmids carrying the alpA gene
CONDITION	Environmental conditions/changes in conditions	Strains carrying a mutation in the crp structural gene fail to ***repress*** ODC and ADC activities in response to **increased cAMP**
RATE	Change of level or rate	marR mutations ***elevated*** inaA expression by **10- to 20-fold** over that of the wild-type.
DESCRIPTIVE-AGENT	Descriptive information about AGENT	It is likely that HyfR ***acts*** as **a formate-dependent regulator** of the hyf operon
DESCRIPTIVE-THEME	Descriptive information about THEME	The FNR protein ***resembles*** **CRP**.
PURPOSE	Purpose/reason for the event occurring	The fusion strains were ***used*** **to study** the regulation of the cysB gene by assaying the fused lacZ gene product

Named Entity Superclasses

NE class	Definition
DNA	Entities chiefly composed of nucleic acids and their structural or positional references. This includes the physical structure of all DNA-based entities and the functional roles associated with regions thereof.
PROTEIN	Entities chiefly composed of amino acids and their positional references. This includes the physical structure and functional roles associated with each type.
EXPERIMENTAL	Both physical and methodological entities, either used, consumed or required for a reaction to take place.
ORGANISMS	Entities representing individuals or collections of living things and their component parts.
PROCESSES	A set of *event* classes used to label biological processes described in text.

Event Annotation example

DNA
A promoter
AGENT

has been identified that

directs
verb

PROCESS
relA gene transcription
THEME

towards

DNA
the pyrG gene
DESTINATION

in

a counterclockwise direction
MANNER

DNA
on the E. Coli chromosome
LOCATION

The extraction process of Event Frame

Input sentence:

Agent Theme

`transfer operon expresses F-like plasmids`

DNA DNA

Syntactic analysis:

`[Agent:NN:DNA] [Verb:VBZ:express] [Theme:NN:DNA]`

Extracted event frame:

`express(Agent=>DNA, Theme=>DNA)`

The BioLexicon: acquired **bio-event frames**

verb	Bio-event frames
activate	Agent#Theme#
	Agent#Theme#Condition#
	Agent#Theme#Location#
	Agent#Theme#Manner#
	Agent#Theme#Source#
	Theme#
	Theme#Condition#

verb	Bio-event frames with NE types
activate	Agent-DNA#Theme-DNA#
	Agent-Organisms#Theme-Protein#
	Agent-Protein#Theme-DNA#

The BioLexicon: Syntax-Semantics Linking (1)

- The starting point:
  - acquired subcategorization frames
  - verbal bio-event frames based on corpus annotation
  - acquired using different techniques and corpora of different size
- Linking concerned 168 verbs for which both syntactic and semantic information was available
- Linking process carried out manually by a linguist
  - Different information types were taken into account, i.e.
    - literature regarding hierarchies of semantic roles and grammatical functions
      - given a thematic role hierarchy (agent>theme ...) and a syntactic functions hierarchy (subject>object ...), the mapping usually proceeds from left to right
    - a list of 'prototypic' syntactic realisations of semantic arguments
    - exploitation of general language repositories of semantic frames containing both syntactic and semantic information (as possible benchmarks)

The BioLexicon: Syntax-Semantics Linking (2)

- Linking process resulted in 668 linked frames
- Different types of mapping were performed:
  - full mapping (239 frames)
    - arity of the subcategorization and bio-event frames is the same (ISO)
  - partial mapping
    1) semantic frame contains more slots than corresponding subcategorization frame (123 frames) (AUG)

       e.g. `AGENT>ARG1#THEME>ARG2#LOCATION>PP-in#`**`CONDITION>0`**
    2) subcategorized slots do not have a semantic counterpart in the corresponding bio-event frame (166 frames) (RED)

       e.g. **`0>ARG1`**`#THEME>ARG2#DESTINATION>PP-into`
    3) a combination of cases 1) and 2) above (140 frames)

       e.g. **`0>ARG1`**`#THEME>ARG2#SOURCE>PP-from#`**`CONDITION>0`**

The BioLexicon: **syntax-semantics linking**

activate	Theme	ARG2	0	ARG1			RED
	Agent	ARG1	Theme	ARG2	Condition	PP-in	ISO
	Agent	ARG1	Theme	ARG2	Manner	PP-by	ISO
	Agent	ARG1	Theme	ARG2	Location	PP-in	ISO
	Agent	ARG1	Theme	ARG2	Source	0	AUG
	Theme	ARG2	Condition	PP-in	0	ARG1	RED
	Agent	ARG1	Theme	ARG2			ISO
	Theme	ARG1					ISO
	Agent	ARG1	Theme	ARG2	Manner	PP-in	ISO

Useful information for **mixed syntax-semantics approaches**

G. Venturi, S. Montemagni, S. Marchi, Y. Sasaki, P. Thompson, J. McNaught, S. Ananiadou, 2009, "Bootstrapping a Verb Lexicon for Biomedical Information Extraction", in Proceedings of the CICLing-2009 conference, Mexico.

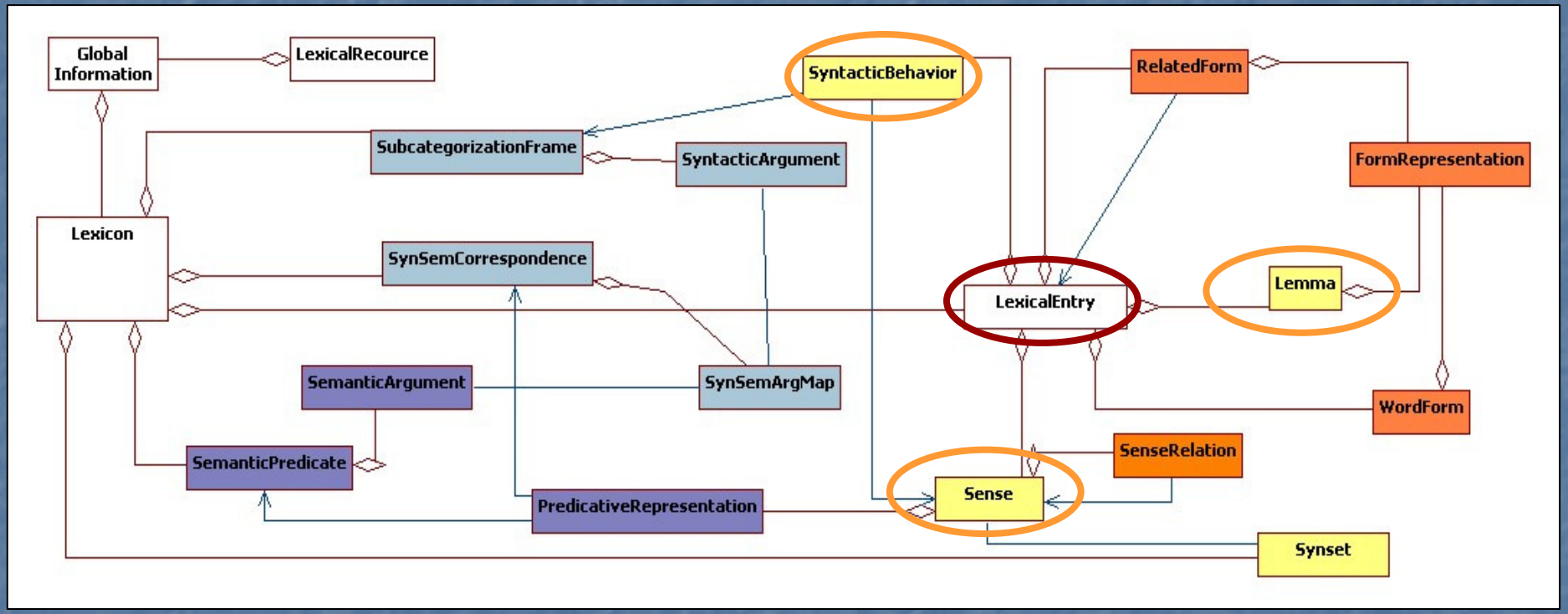
From bricks of biolexical knowledge to a computational BioLexicon

The BioLexicon: **representation model**

- The BL model is conformant to ISO-LMF (ISO 24613:2008)
  - **high-level objects**: the meta-model, i.e. a set of independent lexical objects with relations among them
  - **low-level objects**: a set of Data Categories, i.e. linguistic *constants* in the form of attribute-value pairs (either drawn from the ISO-12620 or defined for the special domain)
- XML DTD for the entire lexicon
- The implementation consists of a flexible, extensible relational MySQL database
- Automatic population procedures relying on a dedicated input data structure, the BioLexicon XML Interchange Format (XIF)
- An XML LMF conformant export function is available

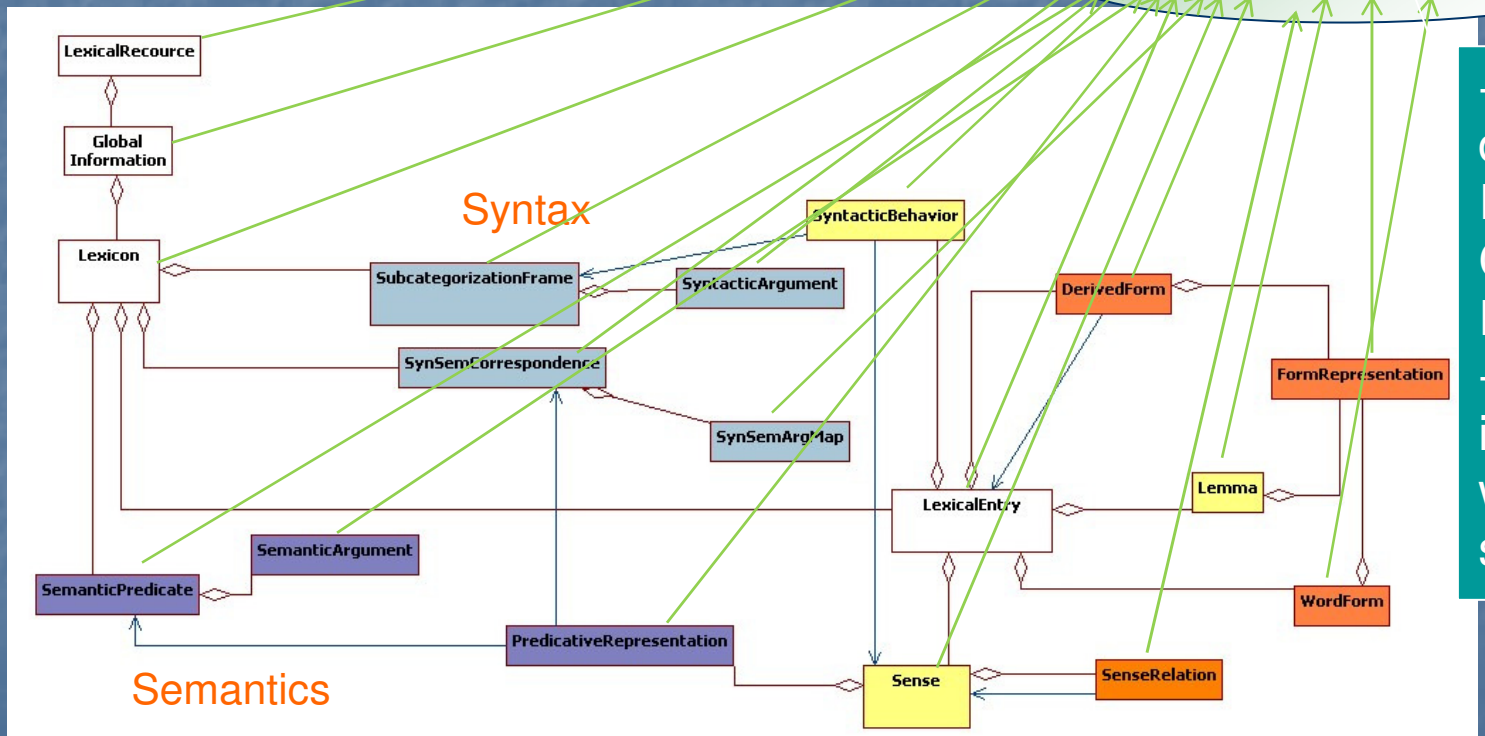
The BioLexicon: **representation model**

The BioLexicon Model: High-level objects, lexical objects



The BioLexicon: **representation model**

BioLexicon Model: Data Categories



- Partially drawn by the ISO Data Category Registry
- Partially integrated with domain-specific DCs

The BioLexicon: **the starting point**

Semantic type	Resources
Cell	Cell ontology
Cell Component	Gene Ontology GO:0005575 cellular component
Chemical	CHEBI, IMR:0000947 chemical
Disease	OMIM
Enzyme	Enzyme commission
Gene	BioThesaurus
Ligand	IMR - INOH Protein name/family name ontology
Nuclear Receptor	GO:0004879 ligand-dependent nuclear receptor activity

Semantic type	Resources
NucleicAcid Region	Sequence Ontology :Region
Operon	RegulonDB, ODB (Operon DataBase)
Organism	NCBI Species
Transcription Factor-BindingSite	Sequence Ontology
Protein	BioThesaurus
Protein Complex	Corum database
Protein Domain	InterPro
Transcription Regulator	RegulonDB, TransFac, Gene Ontology Annotation

The BioLexicon **by numbers**

Entries and variants by semantic type

Sem. Type	# Entries	# Variants
Gene/Prot	**1640608**	**1408312**
Gene/Prot (synsets)	358335	936126
Organisms	**482992**	**182610**
Enzymes	4016	4164
Protein Domains	**16940**	**15412**
Protein Compl.	2104	418
Chemicals	**19637**	**77475**
Diseases	19457	11314
Molecular Roles	**8850**	**29831**
Cell	842	512
Trans. Factors	**160**	**129**
Operons	2672	368
Sequences	**1431**	**741**

Entries by part of speech

POS	# entries
Nouns	**2231574**
Adjectives	3428
Verbs	**1154**
Adverbs	550

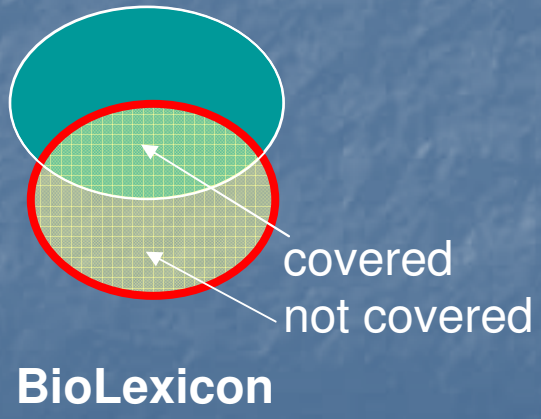
Verbs

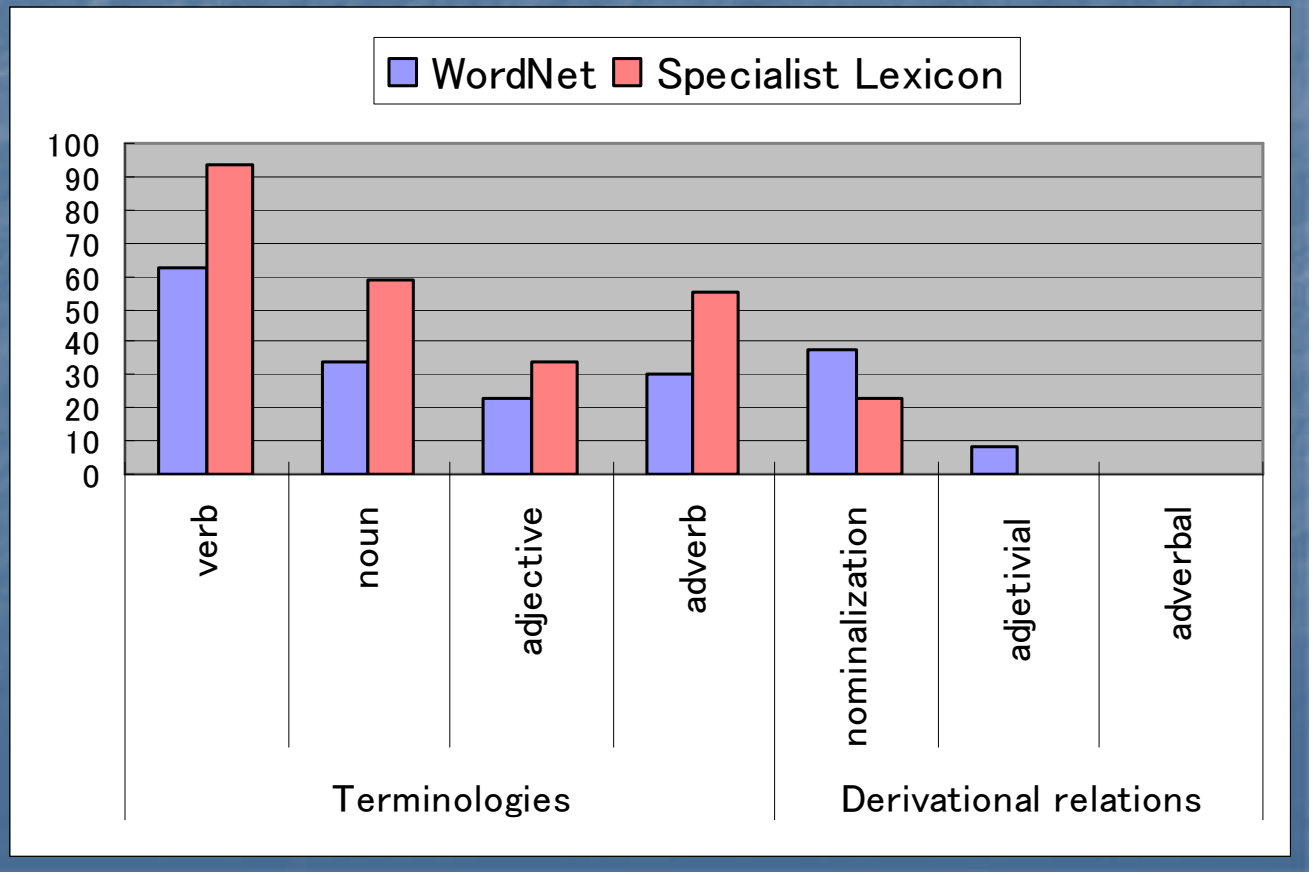
domain-specific	general
658	496

inflected forms	**15274**	
related entries (e.g. absorb -> absorption/N, absorber/N, absorbing/J, absorbable/J, absorbent/J, absorbently/R)	2764	-
verb-SCF associations	**3040**	**-**
verb-SLOT associations	1710	-
bio-event frames	**856**	**-**
syntax-semantics mappings (concerned with 168 verbs)	668	-

The BioLexicon: **intrinsic evaluation**

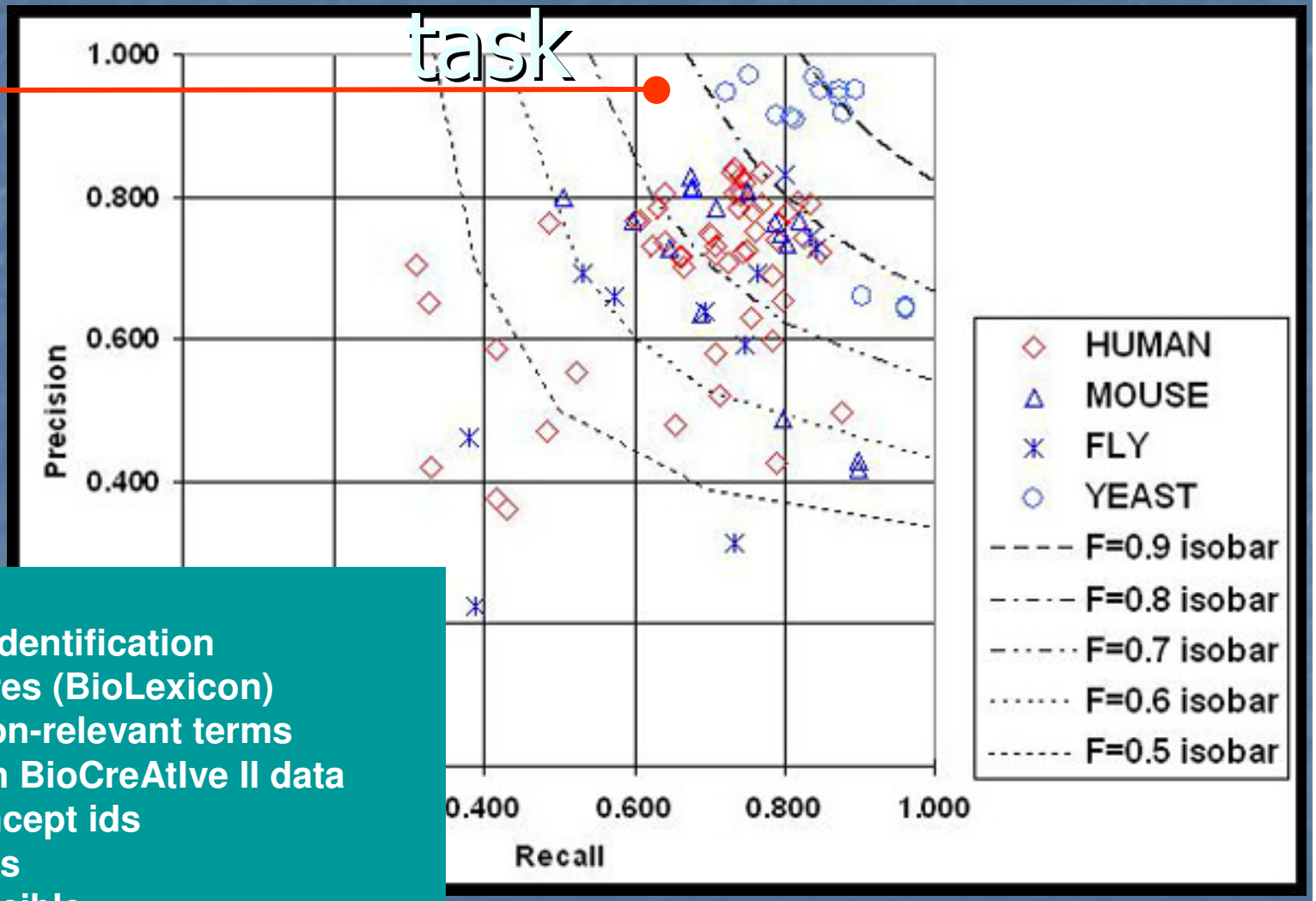
- Comparison with two existing large-scale dictionaries
  - WordNet: General English Thesaurus
  - NLM Specialist Lexicon: Biomedical Lexicon
- Coverage evaluation





BioLexicon in BioCreAtIve II, GN task

Biolexicon, human



Applied methods:
- **Abner for gene identification**
- **Statistical features (BioLexicon) for filtering of non-relevant terms**
- **Classification on BioCreAtIve II data**
- **Only human concept ids**

⇒ Baseline results
⇒ Highly reproducible
⇒ Available as Whatizit module (BioLexHuman)

The BioLexicon: **extrinsic evaluation**

Task-based evaluation (still ongoing)

Task	Data	Tool
IR	TREC Genomics Track 2007	▪BLTagger ▪NeMine (NER)
IE	UoM Gene Regulation Corpus	▪BLTagger ▪NeMine (NER) ▪Enju with the BL

NeMine (http://text0.mib.man.ac.uk/~sasaki/bootstrep/nemine.html)

Y. Sasaki, P. Thompson, J. McNaught, S. Ananiadou, 2009, "Three BIONLP tools powered by the BioLexicon", in Proceedings of EACL 2009.

BioLexicon **distribution**

- The BioLexicon (MySQL version) is distributed through the ***European Language Resources Association*** (**ELRA**)
  - http://www.elra.info or http://www.elda.org

- Benefits
  - Servicing of bug reporting through ELRA
  - Organisational embedding into other lexical resources
  - Long-term availability
  - Support to European language infrastructures

- Different licence types for
  - Commercial use
  - Research use by commercial organisations
  - Research use by academic organisations

Conclusions

A unique resource amongst large-scale computational lexicons within the biomedical domain in terms of **coverage** and **typology of contained information**

Designed to meet **bio-Text Mining requirements**

Populated with info from available **biomedical resources** and **texts**

Semi-automatically populated from **corpora**: *Population toolkit available*

Including both **domain-specific** and **general language** words

Conformant to **international lexical representation standards**

Rich linguistic information ranging over different linguistic descriptions levels

TRY IT!

THANK YOU